第 5 章

智 算 中 心

智算中心是人工智能发展的关键基础设施，是带动科技创新、经济发展、社会治理水平提升及赋能全产业链发展的重要力量。本章首先介绍了算力和智算中心的概念，以及智算中心服务分类，然后讨论智算中心的功能架构和关键技术，最后探讨智算中心的应用场景及高校智算中心具体案例。

5.1 智算中心概述

智算中心作为信息基础设施的重要组成部分，通过算力的生产、聚合、调度和释放，能够为快速增长的人工智能算力需求提供基础支撑，在推进人工智能产业化、赋能产业人工智能化、促进产业集群化等方面具有显著作用，是数字经济时代促进科技创新、优化产业结构、提升国家竞争力的重要支撑。

5.1.1 算力的概念

1. 算力的定义

算力，就是通过对信息数据进行处理，实现目标结果输出的计算能力。简单来说，不论是人类文明初期的口算、心算；还是结绳记事、算盘等工具的应用；或是如今计算机、服务器芯片的高速运转，本质上都是对数据进行运算处理。

算力，即计算能力(Computing Power)，是指计算机系统或设备在单位时间内处理数据和执行计算任务的能力。从技术层面看，算力涉及硬件设备的计算性能、数据处理能力以及软件算法对硬件资源的利用效率等多个方面。例如，在训练一个深度学习模型时，算力决定了模型能够在多长时间内完成对大量训练数据的学习，以及最终模型能否达到较高的准确率。在智算中心场景下，算力主要体现在对海量数据进行高效处理、分析和挖掘，以支持人工智能、大数据分析、科学计算等复杂应用的能力。它不仅取决于硬件设备的性能，如CPU、GPU、FPGA、ASIC 等芯片的计算速度和核心数量等，还与软件算法的优化、系统的架构设计及数据的传输效率等因素密切相关。

算力的作用是完成计算任务。大家都知道，计算机硬件系统的运转及程序软件的执行，是由无数个计算任务支撑的。因此，芯片所提供的算力，就是整个系统正常工作的动力来源。算力支撑了所有的 IT 系统，而 IT 系统支撑了整个社会。从这个角度来说，将算力誉

为社会发展的基石，也不为过。总之，算力是集信息计算力、网络运载力、数据存储力于一体的新型生产力。

2. 算力的分类

(1) 按硬件类型划分

CPU 算力：CPU(Central Processing Unit，中央处理器)是计算机的核心部件，负责执行各种指令和数据处理任务。在人工智能领域，CPU 算力主要用于处理一些通用计算任务，如数据预处理、模型推理中的简单逻辑运算等。CPU 具有强大的通用性和灵活性，能够处理各种类型的数据和指令，但在处理大规模并行计算任务时效率相对较低。

GPU 算力：GPU(Graphics Processing Unit，图形处理器)最初是为图形渲染而设计的，具有大量的计算核心和强大的并行计算能力。在人工智能中，GPU 算力被广泛应用于深度学习模型的训练和推理。GPU 能够同时处理多个数据，加速了矩阵运算和卷积运算等深度学习算法中的关键操作。

FPGA 算力：FPGA(Field-Programmable Gate Array，现场可编程门阵列)是一种可编程的硬件设备，用户可以根据自己的需求对其内部逻辑进行编程。在人工智能领域，FPGA 算力常用于对实时性要求较高、功耗要求较低的应用场景。FPGA 具有低延迟、高能效比的特点，可以根据特定的算法进行定制化设计，实现高效的计算。

ASIC 算力：ASIC(Application-Specific Integrated Circuit，专用集成电路)是为特定应用而设计的集成电路，具有高度的专业化和定制化特点。在人工智能中，ASIC 算力通常用于大规模的商业应用，如数据中心和云计算平台等。ASIC 具有很高的计算性能和能效比，但开发成本高、周期长，且灵活性较差。如谷歌的 TPU(Tensor Processing Unit，张量处理器)就是一种专门为深度学习设计的 ASIC 芯片，能够为谷歌的搜索和语音识别等服务提供强大的算力支持。

(2) 按计算类型分类

通用算力：基于 CPU 提供的计算能力，适用于各种通用计算任务，如办公自动化、数据库管理、Web 服务等。CPU 具有强大的通用性和灵活性，能够处理各种类型的数据和指令。通用算力在处理复杂逻辑和顺序任务时表现出色，但在处理大规模并行计算任务时效率相对较低。

智能算力：主要由 GPU、FPGA、ASIC 等芯片提供，专门用于人工智能领域的计算任务，如深度学习模型的训练和推理。这些芯片具有高度的并行计算能力，能够加速矩阵运算和卷积运算等人工智能算法中的关键操作。智能算力在处理大规模数据并行计算任务时具有显著优势。

超算算力：由超级计算机提供的强大计算能力，主要用于科学计算、工程模拟、气象预报等对计算性能要求极高的领域。超级计算机通常由大量的计算节点组成，采用高速互联网络进行连接，能够实现大规模的并行计算。超算算力具有极高的计算精度和强大的并行处理能力，能够处理复杂的科学模型和大规模的数据集。

(3) 按算力规模分类

小算力：小于 20000 TFLOPS。

中算力：20000～80000 TFLOPS。

大算力：大于 80000 TFLOPS。

提示：FLOPS(Floating-Point Operations Per Second)是用于衡量计算机浮点运算能力的指标。浮点运算在科学计算、工程模拟、图形处理等领域非常重要。

以下是 TFLOPS 与其他常见 FLOPS 单位的换算关系。

1 FLOPS = 每秒 1 次浮点运算。

1 KFLOPS = 10^3 FLOPS = 每秒 1000 次浮点运算。

1 MFLOPS = 10^6 FLOPS = 每秒 100 万次浮点运算。

1 GFLOPS = 10^9 FLOPS = 每秒 10 亿次浮点运算。

1 TFLOPS = 10^{12} FLOPS = 每秒 1 万亿次浮点运算。

1 PFLOPS = 10^{15} FLOPS = 每秒 1000 万亿次浮点运算。

1 EFLOPS = 10^{18} FLOPS = 每秒 100 万亿亿次浮点运算。

【例 5-1】 某科研机构正在评估其新购置的超级计算机的算力规模，以便合理分配计算资源。该超级计算机的浮点运算能力为 65 000 TFLOPS。

① 请将该超级计算机的算力转换为 GFLOPS 和 PFLOPS 单位表示。

② 根据给定的算力规模分类标准(小算力：小于 20 000 TFLOPS；中算力：20 000～80 000 TFLOPS；大算力：大于 80 000 TFLOPS)，判断该超级计算机属于哪一类算力规模，并简要说明理由。

【解】

① 单位换算

将 65 000 TFLOPS 转换为 GFLOPS。

已知 1 TFLOPS = 10^{12} FLOPS，1 GFLOPS = 10^9 FLOPS

那么将 TFLOPS 换算为 GFLOPS，需要乘以 10^3(因为 $10^{12} \div 10^9 = 10^3$)。

所以 65 000 TFLOPS 换算为 GFLOPS 为：

$$65\ 000 \times 10^3 = 65000000\ \text{GFLOPS} = 6.5 \times 10^7\ \text{GFLOPS}$$

将 65 000 TFLOPS 转换为 PFLOPS

因为 1 PFLOPS = 10^{15} FLOPS，1 TFLOPS = 10^{12} FLOPS

所以将 TFLOPS 换算为 PFLOPS，需要除以 10^3

(因为 $10^{12} \div 10^{15} = 10^{-3}$，也就是乘以 10^{-3})

则 65 000 TFLOPS 换算为 PFLOPS 为：$65000 \times 10^{-3} = 65$ PFLOPS。

② 算力规模分类判断

根据给定的算力规模分类标准：

小算力：小于 20 000 TFLOPS

中算力：20 000～80 000 TFLOPS

大算力：大于 80 000 TFLOPS

由于该超级计算机的算力为 65 000 TFLOPS，20 000 < 65 000 < 80 000，满足中算力的范围，因此该超级计算机属于中算力规模。

5.1.2　智算中心的定义

智算，即智能算力，是基于 CPU 与 GPU 等加速芯片异构组合、支撑人工智能创新应用的一种高性能算力服务。作为一种虚拟计算能力，智能算力在现实世界需要物理实体承载

支撑，包括智能终端、边缘设备和智算中心等多种承载形态。伴随以人工智能大模型为代表的人工智能技术加速演进，对规模化、高性能训推算力集群提出要求，智算中心成为智能算力的主流物理载体，受到社会各方关注。

智算中心，即人工智能计算中心（Artificial Intelligence Data Center，AIDC），一般认为是在传统数据中心的基础上，基于 GPU、TPU、FPGA 等人工智能芯片及计算框架构建的人工智能基础设施，可以支撑大量数据处理和复杂模型训练。狭义来看，智算中心是智能算力的物理载体，是“机房＋网络＋GPU 服务器＋算力调度平台”的融合基础设施，是传统数据中心的增值性延伸。广义来看，智算中心是人工智能软硬件技术一体化的载体，是“算力＋数据＋算法”的融合服务，是促进人工智能产业化和产业人工智能化的重要引擎，是传统云的智能化升级。

智算中心是指具备智能计算、存储、高性能网络、容器和安全等基础设施和服务，通过机器学习平台和大模型智算平台为各类智算场景和应用提供人工智能算力、大模型开发训练和统一监控运营等服务的系统。

智算中心不仅应具备通用数据中心的计算、存储和网络等基础功能，还应提供强大的计算资源，满足日益丰富的人工智能应用开发与场景试验需求。它通过算力的生产、聚合、调度和释放，驱动人工智能模型对数据进行深度挖掘和加工，支撑数据开放共享，从而促进人工智能技术产业化，强力推动算力生态的可持续发展，成为支撑经济数智化和社会数智化的重要引擎。

与传统的数据中心有所不同，智算中心并非简单的数据存储仓库，而是聚焦于人工智能领域的计算需求。它集成了众多高性能的计算设备，如 GPU 和 TPU 等，这些芯片就像是工厂里的“超级工人”，擅长并行计算，能够快速处理复杂的数学运算，大幅提升计算效率。与此同时，高速网络连接如同工厂内部的“高速传送带”，确保数据在各个设备之间快速、流畅地传输，让计算过程无缝衔接。

此外，智算中心还配备了专业的软件系统，犹如工厂的“智能管理系统”，涵盖人工智能框架、资源管理软件及数据管理软件等，它们能够简化模型开发流程，自动调度计算资源，并保障数据的质量与安全。

5.1.3 智算中心的功能

智算中心的两大核心功能是数据处理能力和推理服务。数据处理关乎模型训练的质量和速度，推理服务则直接影响了业务逻辑的实时响应和执行效率。

智算中心的主要功能包括以下 4 个方面。

（1）数据存储与分析

智算中心具备大规模、高可靠性的数据存储能力，能够安全、可靠地存储海量数据，包括结构化数据和非结构化数据，并利用强大的计算力，对数据进行快速处理和分析，提取有价值的信息。

（2）人工智能模型训练与优化

智算中心为人工智能模型的训练提供了强大的计算支持，加速模型的收敛速度，提高训练效率。此外，通过对训练好的模型进行优化，能够提高模型的准确性和泛化能力。

(3) 应用开发与创新支持

智算中心为开发者提供了丰富的开发平台和工具,支持他们进行各种应用的开发,包括人工智能应用和大数据应用等,科研机构、高校、企业等可在智算中心的基础上进行创新应用的探索和孵化。

(4) 算力服务与共享

智算中心将算力资源以服务的形式提供给外部用户,包括科研机构、高校、企业等,满足他们对算力的临时或长期需求,进而实现算力资源在不同用户和应用之间的共享,提高算力资源的利用率。

5.1.4　智算中心服务分类

从智算中心产品功能和算力规模角度,智算中心为用户提供的服务场景可分为简单智算服务、中等智算服务和大模型智算服务。

1. 简单智算服务

简单智算服务为用户提供通用的 CPU 和 GPU 等算力,用户可通过 API 调用 NLP 等通用模型获取推理服务,支持针对不同数据规模提供小算力、中算力和大算力的推理服务,适用于图像分类、内容推荐和生产过程统计等典型场景。

2. 中等智算服务

基于通用模型,支持针对不同数据规模提供小算力、中算力和大算力的训练和推理等算力服务,适用于文本分类、命名实体识别、事件提取、错别字检查、制造质量分析、医疗数据分析等各领域算力服务场景。

3. 大模型智算服务

基于大语言模型(模型参数量在 10 亿以上),为用户提供通用模型预训练(大算力)、行业模型预训练(中算力)、场景微调训练(小算力)等,适用了 AIGC 典型场景。

(1) 医疗医药文献摘要提取

针对医学部门提交的三类药物警戒临床文献,准确识别药物警戒的五要素(可识别的上报人,可识别的患者,怀疑药物,不良反应,相关性描述),抽取并总结生成完整的综述报告,人工校准后翻译为对应的英文报告。

(2) 智能金融分析服务

针对金融行业,基于大模型智算服务,可以提供金融市场的深度分析。例如,通过分析海量的金融新闻、财报和市场数据等,大模型能够识别市场趋势、预测股票走势、评估投资风险,并为投资者提供个性化的投资建议。同时,大模型还可以用于金融欺诈检测,通过分析交易模式和行为数据,及时发现并防范潜在的欺诈行为。

(3) 智能媒体内容创作

针对媒体行业,大模型智算服务可以支持智能媒体内容创作。例如,通过训练大模型掌握新闻写作、故事创作、广告文案生成等技能,媒体工作者可以输入关键词或主题,大模型能够自动生成高质量的新闻稿件、故事情节或广告文案。同时,大模型还可以用于视频内容的智能剪辑和配音,根据视频内容和风格自动选择合适的音乐、字幕和配音,提升媒体内容的制作效率和观赏性。

5.1.5 智算中心与数据中心、超算中心

智算中心、数据中心和超算中心作为信息技术领域的三大基础设施，各自承担着不同的角色和功能。

智算中心是以人工智能计算为核心，整合 GPU、FPGA 等专用硬件，构建面向人工智能模型训练和推理的专用计算平台。其核心功能是提供高效的人工智能算力支持，加速深度学习、自然语言处理等人工智能应用的开发部署，推动人工智能技术的产业化落地。

数据中心是基于云计算技术构建的 IT 基础设施，通过虚拟化技术整合计算、存储和网络资源，以服务形式对外提供。其核心功能是实现资源的弹性分配和按需使用，支持企业数字化转型、互联网应用部署等场景，强调灵活性和可扩展性。

超算中心是配备高性能计算（High Performance Computing，HPC）集群的设施，专注于解决科学计算和工程模拟等需要极高算力的复杂问题。其核心功能是提供大规模并行计算能力，支持气象预报、基因测序、航空航天设计等对计算精度和实时性要求极高的领域。

智算中心与数据中心、超算中心三者之间的区别如表 5-1 所示。

表 5-1 智算中心、数据中心和超算中心的区别

主要指标	智 算 中 心	数 据 中 心	超 算 中 心
建设目的	促进人工智能产业化、产业人工智能化、政府治理智能化	帮助企业降本增效或提升盈利水平	面向科研人员和科学计算场景提供支撑
技术标准	统一标准、统筹规划，开放建设、互联互通互操作，高安全标准	标准不一、重复建设，安全水平参差不齐	服务采用并行架构，标准不一，存在多个技术路线，互联互通难度较大
具体功能	算力生产供应平台、数据开放共享平台、智能生态建设平台、产业创新聚集平台	能以更低成本承载企业和政府等用户个性化、规模化业务应用需求	以提升国家及地方自主科研创新能力为目的，重点支持各种大规模科学计算和工程计算任务
应用领域	面向人工智能典型应用场景，赋能各行各业，如自然语言处理、智能制造、自动驾驶、智慧农业等	面向众多应用场景，应用领域和应用层级不断扩张，支撑构造不同类型的应用	基础学科研究、工业制造、生命医疗、模拟仿真、气象环境、天文地理等
投-建-运模式	政府主导下的政企合作共建模式，政府出资指导建设，企业承建运营	行业巨头或政府投资建设，其他用户按需付费使用，以数据服务盈利	政府科研单位投资建设、运营

5.2 智算中心的功能架构

智算中心的功能架构如图 5-1 所示，它是一个综合性的系统，由智算基础设施、智算基础平台、智算服务平台和统一监控运营平台四大核心部分组成，共同支撑简单智算场景、中等智算场景及大模型智算场景的应用与发展。这一架构不仅体现了技术的先进性，也确保了系统的灵活性与可扩展性，以适应不断变化的计算需求。

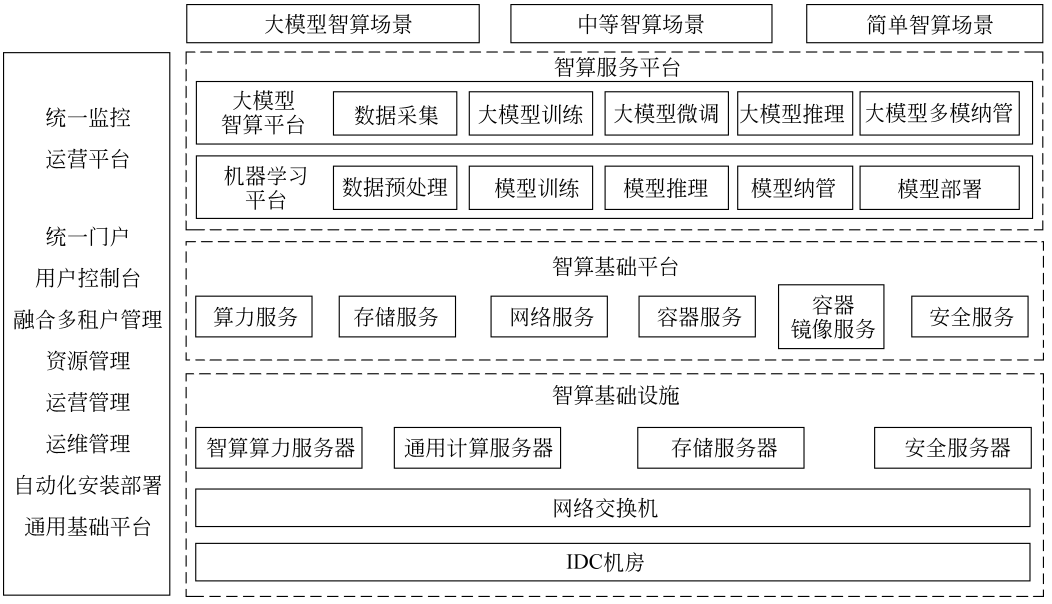

图 5-1　智算中心功能架构

5.2.1　智算基础设施

智算基础设施是智算中心的物理基石，它为整个系统提供了稳定可靠的运行环境。这一部分主要包括了 IDC(Internet Data Center，互联网数据中心)机房、网络交换机、智算算力服务器、通用计算服务器、存储服务器及安全服务器等关键设备，如图 5-2 所示。

图 5-2　智能基础设施

IDC 机房：作为数据存储与处理的物理空间，IDC 机房不仅提供安全可靠的运行环境，还集成先进的温控系统确保温度恒定，采用冗余供电设计保障电力不间断，并配备智能消防系统应对突发情况。同时，机房实施严格的访问控制与 24h 监控，确保设备在最佳状态下稳定运行。

网络交换机：作为智算中心的数据传输枢纽，采用高速、低延迟设计，支持 InfiniBand 及 RoCEV2 等高性能网络架构，确保多节点间数据高效流通。交换机配备冗余电源与智能管理软件，实现网络故障快速恢复与动态路由，为大规模数据处理提供可靠的网络设施。

智算算力服务器：专为大模型智算场景设计，搭载高性能 GPU/TPU 加速器，支持深

度学习和大数据分析等高强度计算任务。其支持多框架并行计算，配备高速数据通道与大容量内存，结合虚拟化技术实现算力调度，满足大规模模型训练与推理需求，是智算中心的核心算力支撑。

通用计算服务器：具备多核 CPU 与 DDR4 内存，支持虚拟化技术，可灵活分配计算资源。适用于轻量级应用至中等规模计算任务，提供快速部署能力。其支持图像预处理和自然语言处理等场景推理，并配备多网络接口与存储协议，满足多样化业务需求，是智算中心的基础计算单元。

存储服务器：采用了分布式存储架构，提供了高容量、高可靠性的数据存储服务，支持大规模数据的快速读写与备份恢复，确保了数据的安全性与可访问性，是智算中心数据持久化的基础设施。

安全服务器：通过集成硬件可信执行环境、防火墙、入侵检测系统及数据加密技术，构建多层次防护体系。支持应用控制、Web 防护、漏洞扫描等功能，并配备负载均衡与态势感知模块，实时监控安全事件。实现访问控制、漏洞管理及多账号统一安全防护，确保智算中心的数据安全。

5.2.2 智算基础平台

智算基础平台依托于智算基础设施，为上层应用提供了一系列核心服务。这些服务包括但不限于算力服务、存储服务、网络服务、容器服务、容器镜像服务及安全服务。

算力服务：支持通用算力和智算算力两种计算方式，满足不同应用场景需要。通过弹性伸缩满足轻量级到大规模并行计算需求。提供批处理和实时流处理等多任务调度方式，支持多种操作系统，结合高可用性设计确保持续稳定运行。用户可根据业务需求灵活配置资源，实现计算效率最大化。

存储服务：提供高性能存储能力，涵盖文件、对象、块存储等多种方式，支持数据备份与恢复。采用数据压缩和加密技术保障安全，实现容器化存储与业务稳定性保障。支持存算分离架构与多种存储介质选择，提供低延迟检索与并行文件存储，满足大规模数据高效管理需求。

网络服务：采用高性能 RDMA(Remote Direct Memory Access，远程直接内存访问)网络(如 InfiniBand、RoCE)，优化数据传输带宽与延迟，支持超大规模网络架构。通过 SDN(Software Defined Network，软件定义网络)技术实现虚拟专用云网络、流量智能调度及负载均衡，确保高可用性和高吞吐量。

容器服务：基于 Docker、Kubernetes 等技术，提供轻量级、可移植和可扩展的容器环境，支持应用快速部署与隔离运行。通过容器编排机制实现高可靠性和高可用性，支持容器自动发现与扩展，并强化网络隔离与安全保护。

容器镜像服务：提供灵活的容器镜像管理与分发系统，支持常见的镜像仓库，如 Docker Hub 及私有仓库。通过高效构建、安全存储及自动化标签管理，确保镜像版本的一致性。

安全服务：为智算平台提供多层次防护，包括严格的身份认证与细粒度访问控制，确保仅授权用户可访问资源。采用数据加密技术保障传输与存储安全，支持态势感知与实时监控，自动检测并处理安全事件。同时提供 Web 防护、负载均衡及漏洞管理，全方位保障数据与应用安全。

5.2.3　智算服务平台

智算服务平台作为连接基础平台与实际应用的桥梁，其重要性不言而喻。该平台主要包括机器学习平台和大模型智算平台两大核心部分，为不同规模与复杂度的智算任务提供了全面的支持。

1. 机器学习平台

机器学习平台如图 5-3 所示，其业务流程包括数据预处理、模型训练、模型纳管和模型部署 4 个阶段。数据源经过数据清洗、数据标注、特征工程和数据增强等预处理后，进行模型的开发。开发完成的模型进行分布式训练、模型推理和模型评估。评估达到要求的，进行下一步的模型纳管；评估未达到要求的，进行模型优化和算法优化后，重新进行分布式训练和模型推理。模型纳管将评估通过的模型进行模型注册、封装、加速和发布，模型仓库对模型进行统一存储和管理。发布后的模型进入模型部署，包括环境配置、自动化部署和版本控制等。

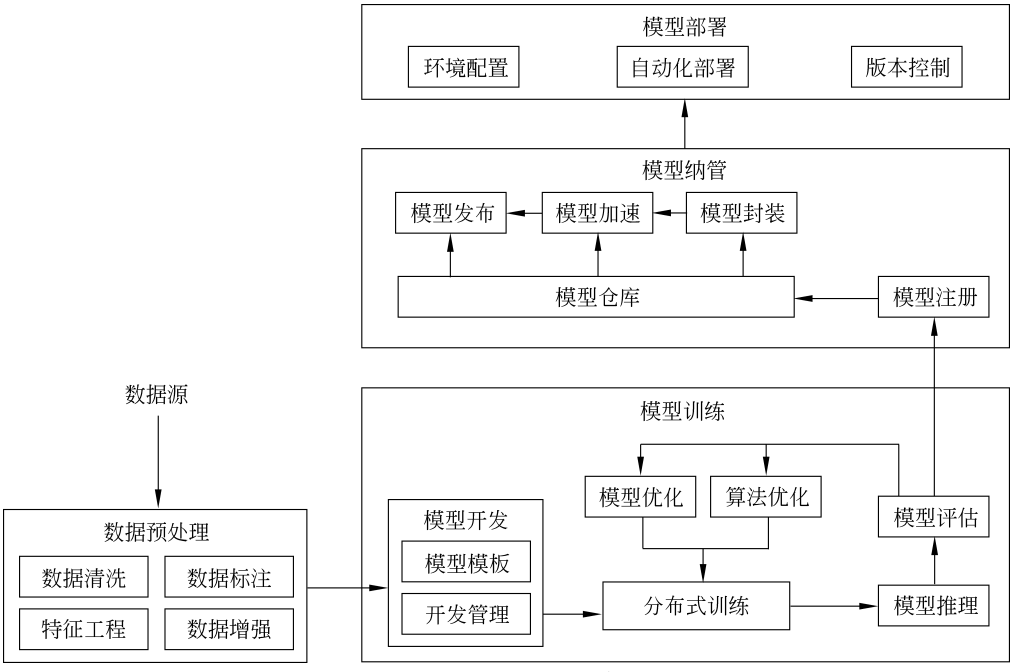

图 5-3　机器学习平台

2. 大模型智算平台

大模型智算平台如图 5-4 所示，其业务流程包括数据采集、数据预处理、大模型训练、大模型多模纳管和模型部署 5 个阶段。数据源首先经过数据采集后进行数据预处理和模型的开发。开发完成的模型进行大模型训练、大模型微调、大模型推理和模型评估。评估达到要求的，进行下一步的大模型多模纳管；评估未达到要求的，进行模型优化和算法优化后，重新进行大模型训练、大模型微调和大模型推理。模型纳管将评估通过的模型进行模型注册、封装、加速、发布和管理，之后完成模型的部署。

在实际应用中，可根据用户需求，提供下列 3 种能力。

- 通用大模型预训练能力。
- 行业大模型预训练和微调能力。
- 针对行业具体场景的微调能力。

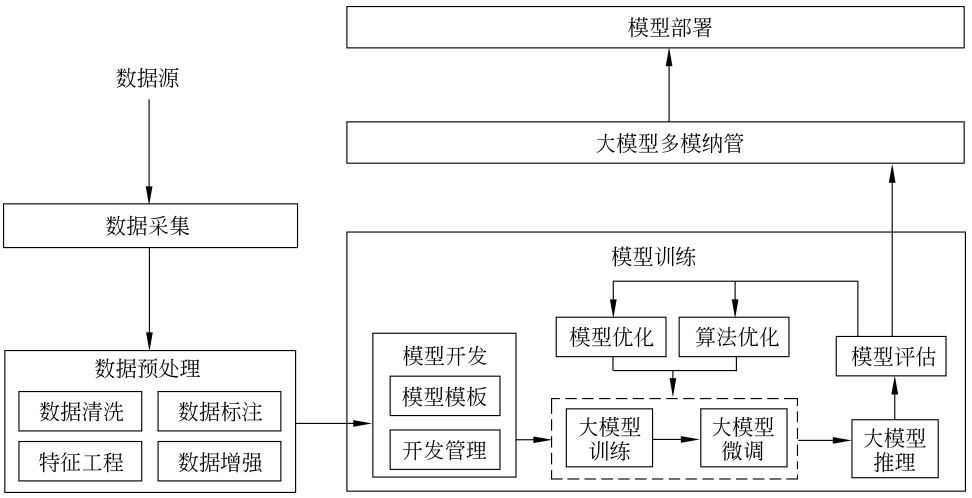

图 5-4 大模型智算平台

5.2.4 统一监控运营平台

统一监控运营平台作为智算中心的管理中枢，发挥着至关重要的作用。它集成了全面的监控、管理和运维能力，确保智算中心能够高效、稳定地运行。该平台由多个关键部分组成，每个部分都承担着特定的功能，共同协作以支持智算中心的日常运营和管理。

统一门户：用户与智算中心交互的窗口，它为用户提供了一个统一的访问入口。通过统一门户，用户可以快速获取所需要的信息和服务，不必在多个系统之间切换。同时，统一门户还提供了个性化的操作界面和自定义配置选项，使用户能够根据自己的需求和使用习惯进行界面布局和功能配置，提升了用户体验和工作效率。

用户控制台：用户进行日常操作的主要界面。它支持用户进行资源管理、任务提交和结果查看等操作，提供了直观、易用的操作界面和丰富的功能选项。用户可以通过用户控制台轻松地管理自己的资源，提交计算任务，并实时查看任务进度和结果，简化了操作流程，提高了工作效率。

融合多租户管理：它是统一监控运营平台的重要特性之一，支持多租户共享资源，同时通过资源隔离和访问控制技术，确保了各租户间的数据隔离和安全性。不同租户之间的数据和操作互不干扰，保证了每个租户都能够在一个安全、独立的环境中运行自己的应用和服务。

资源管理：负责对计算、存储、网络等资源进行动态分配和调度。它根据用户的需求和系统的负载情况，智能地分配资源，提高了资源利用率和灵活性。同时，资源管理模块还提供了资源监控和报警功能，能够实时监测资源的运行状态，并在出现异常时及时发出报警，确保资源的稳定运行。

此外，统一监控运营平台还包括运营管理、运维管理、自动化安装部署和通用基础平台等，这里不再进行逐一介绍。